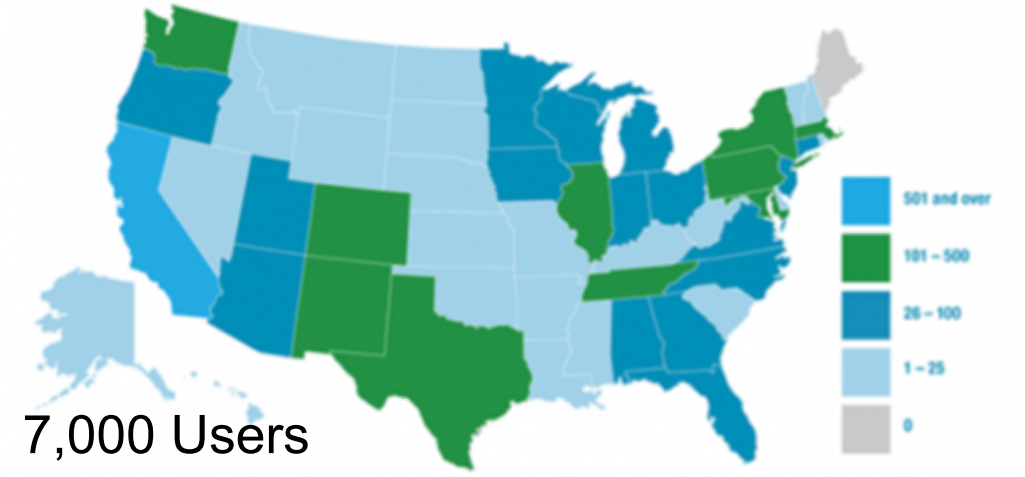
Upcoming computing facilities for science

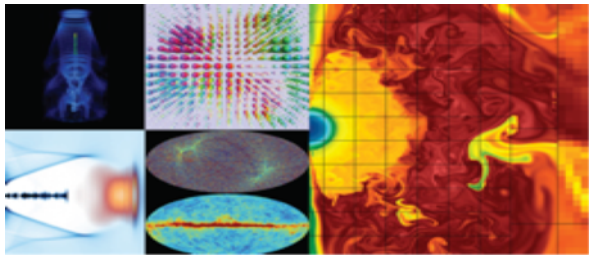
Snowmass Computational Frontier Workshop

Nicholas J Wright
NERSC Chief Architect
11 August, 2020

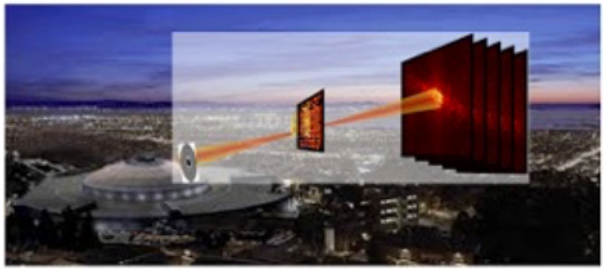
NERSC is the mission High Performance Computing facility for the DOE SC



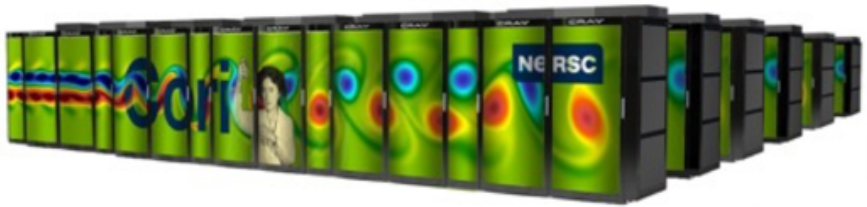
7,000 Users
800 Projects
700 Codes
2000 NERSC citations per year



Simulations at scale



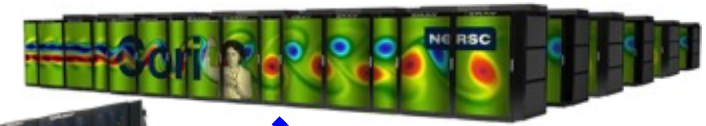
Data analysis support for DOE's experimental and observational facilities
Photo Credit: CAMERA



NERSC Systems Roadmap

NERSC-7: Edison Multicore CPU

2013



NERSC-8: Cori Manycore CPU

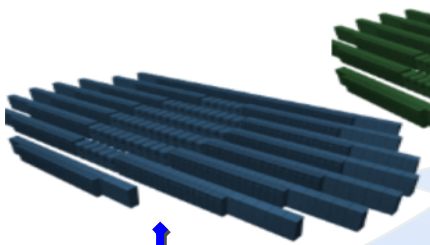
NESAP Launched: transition applications to advanced architectures

2016

NERSC-9: CPU and GPU nodes

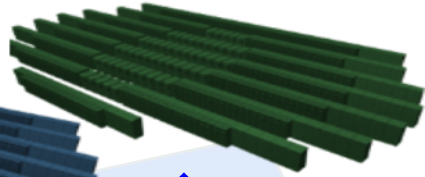
Continued transition of applications and support for complex workflows

2020



NERSC-10: Exa system

2024-5

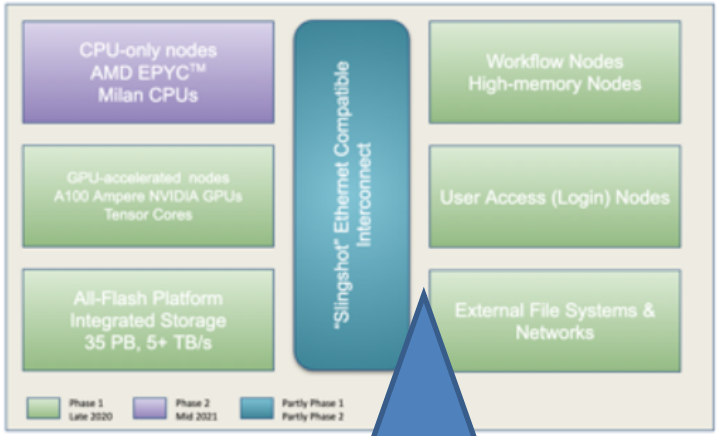


NERSC-11: Beyond Moore

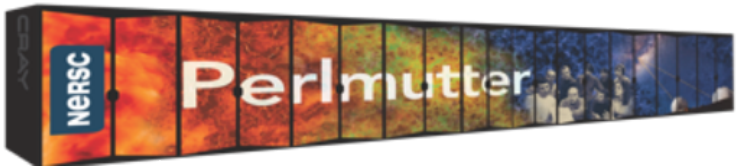
2028

Perlmutter: A System Optimized for Science

- Cray Shasta System providing 3-4x capability of Cori
- GPU-accelerated and CPU-only for large scale simulation and data analysis from experimental facilities
- GPU nodes: 4 NVIDIA A100 “Ampere” GPUs each w/Tensor Cores & NVLink-3 and High-BW memory + 1 AMD “Milan” CPU
  - Over 6000 NVIDIA Volta-Next GPUs
  - Unified Virtual Memory support improves programmability
- Cray “Slingshot” - High-performance, scalable, low-latency Ethernet- compatible network
  - Capable of Terabit connections to/from the system
- Single-tier All-Flash Lustre based HPC file system



Phased delivery
1st phase: End CY2020
2nd phase: Spring CY2021



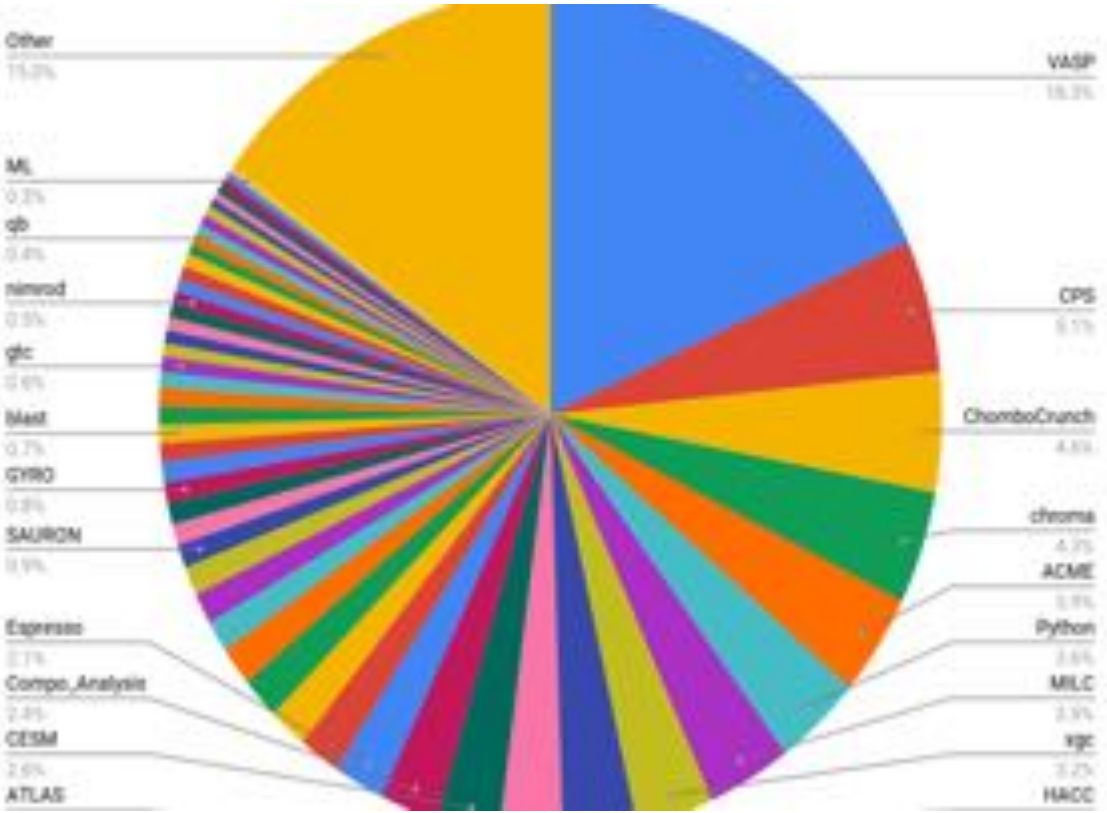
NERSC

BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY

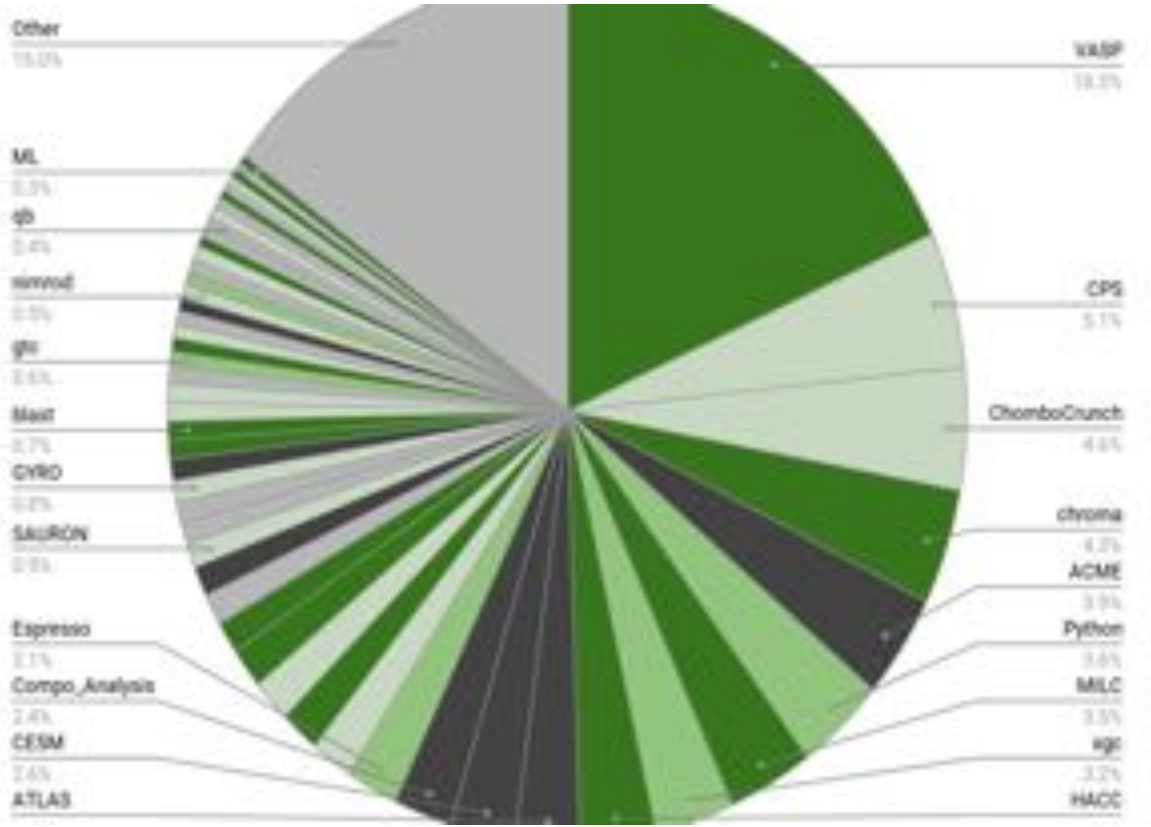
Office of Science

NERSC System Utilization (Aug'17 - Jul'18)



- 3 codes > 25% of the workload
- 10 codes > 50% of the workload
- 35 codes > 75% of the workload
- Over 600 codes comprise the remaining 25% of the workload.

GPU Readiness Among NERSC Codes (Aug'17 - Jul'18)



GPU Status & Description	Fraction
Enabled: Most features are ported and performant	37%
Kernels: Ports of some kernels have been documented.	10%
Proxy: Kernels in related codes have been ported	20%
Unlikely: A GPU port would require major effort.	13%
Unknown: GPU readiness cannot be assessed at this time.	20%

A number of applications in NERSC workload are GPU enabled already.

Pre-Exascale Systems				Exascale Systems
2013	2016	2018	2020	2021-2023

Argonne
Intel/Cray

LBNL
Cray/NVIDIA/AMD

ORNL
Cray/AMD

LANL/SNL
TBD

LLNL
Cray/AMD

Compute Node
2 Intel Xeon scalable "Sapphire Rapids" processors; 6 Xe arch-based GPUs; Unified Memory Architecture; 8 fabric endpoints

Software Stack
Cray Shasta software stack + Intel enhancements + data and learning

GPU Architecture
Xe arch-based "Ponte Vecchio" GPU; Tile-based chiplets, HBM stack, Foveros 3D integration, 7nm

CPU-GPU Interconnect
CPU-GPU: PCIe; GPU-GPU: Xe Link

System Interconnect
System Interconnect Cray Slingshot; Dragonfly topology with adaptive routing

Network Switch
25.6 Tb/s per switch, from 64–200 Gbs ports (25 GB/s per direction)

Sustained Performance
≥ 1 Exaflop DP

High-Performance Storage
≥ 230 PB, ≥ 25 TB/s (DAOS)

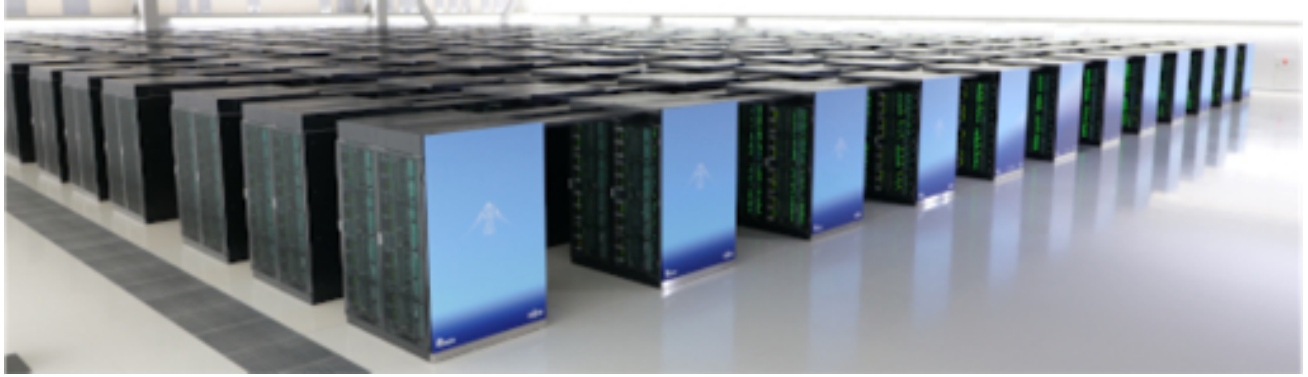
Programming Models
Intel oneAPI, MPI, OpenMP, C/C++, Fortran, SYCL/DPC++

Platform
Cray Shasta

Aggregate System Memory
> 10 PB

Peak Performance	>1.5 EF
Footprint	> 100 cabinets
Node	1 HPC and AI Optimized AMD EPYC CPU 4 Purpose Built AMD Radeon Instinct GPU
CPU-GPU Interconnect	AMD Infinity Fabric Coherent memory across the node
System Interconnect	Multiple Slingshot NICs providing 100 GB/s network bandwidth Slingshot dragonfly network which provides adaptive routing, congestion management and quality of service.
Storage	2-4x performance and capacity of Summit's I/O subsystem. Frontier will have near node storage like Summit.

Fugaku



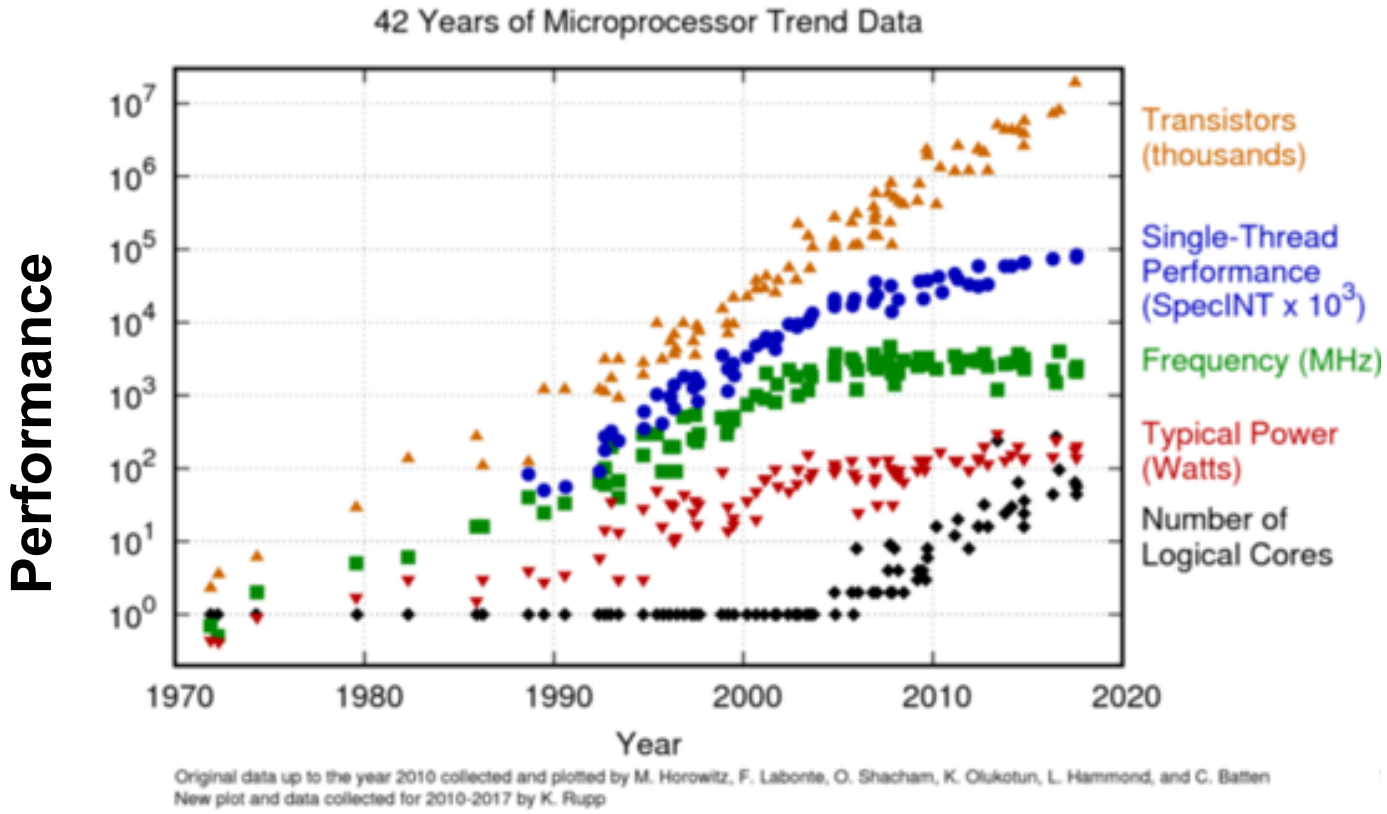
Peak Performance	488 PF
Footprint	158,976 nodes, 414 racks
Node	48 core ARM – 3 TF 32 GiB HBM 1,024 GB/s
System Interconnect	5D TOFU
Storage	Every 16 Nodes local SSD 1.6 TB 150 PB Lustre
Power	29 MW

Further out……

- Seems reasonable to assume all 3 DOE ASCR facilities will upgrade in the mid 20's
  - ~10's EF
- NSF
  - Frontera @ TACC ~ $60M ~40 PF
    - 8,000 dual socket Xenon "Cascade Lake" 28 cores / socket, HDR-IB, 90 4-way NVIDIA GPU Nodes (Volta)
    - Phase II – 10x phase I (~400 PF) due 2023-24

What will these machines look like ?

Technology Scaling Trends



~~> serial performance~~

> Performance per socket

- More cores / CUDA cores
- Multichip packages to keep costs down – NUMA

More heterogeneity

- CPU's and GPU's are here to stay
- AI accelerators ?
- Compute in network / storage

End of Moore's Law ?


DOI:10.1145/3282307

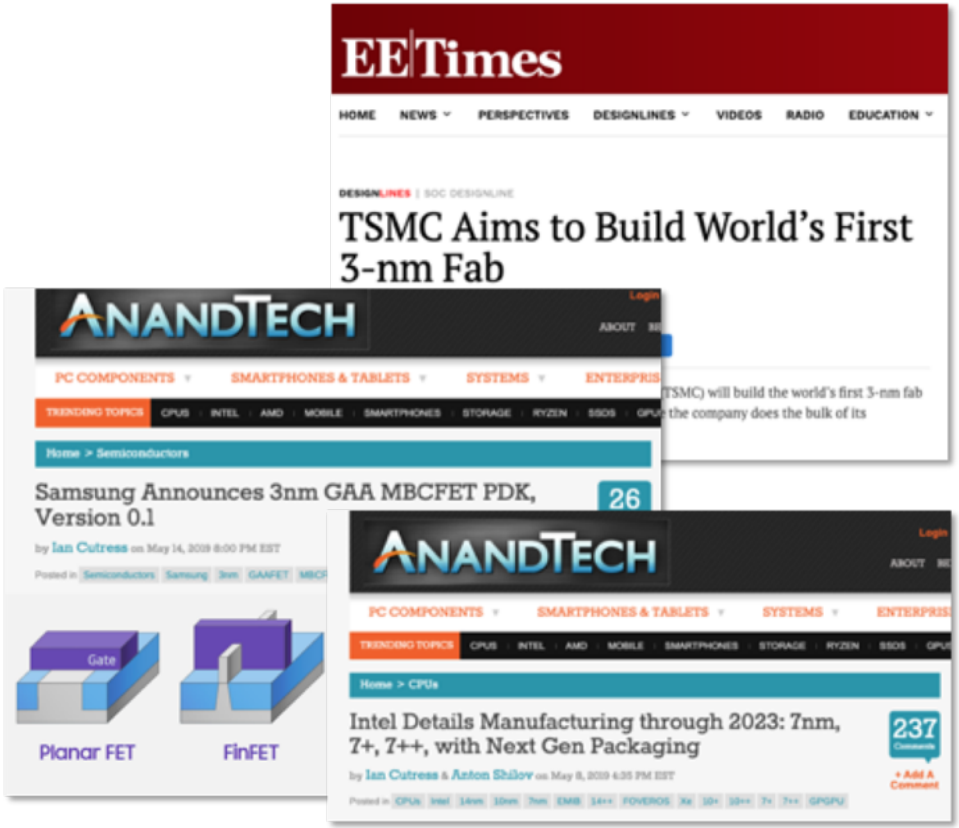
Innovations like domain-specific hardware, enhanced security, open instruction sets, and agile chip development will lead the way.

BY JOHN L. HENNESSY AND DAVID A. PATTERSON

A New Golden Age for Computer Architecture


Extreme Heterogeneity 2018

PRODUCTIVE COMPUTATIONAL SCIENCE IN THE ERA OF EXTREME HETEROGENEITY


EE Times

HOME NEWS PERSPECTIVES DESIGNLINES VIDEOS RADIO EDUCATION

DESIGNLINES | SOC DESIGNLINE

TSMC Aims to Build World's First 3-nm Fab

ANANDTECH

PC COMPONENTS SMARTPHONES & TABLETS SYSTEMS ENTERPRISE

TRENDING TOPICS CPUS INTEL AMD MOBILE SMARTPHONES STORAGE RYZEN SSDS GPUS

Home > Semiconductors

Samsung Announces 3nm GAA MBCFET PDK, Version 0.1

26

by Ian Cutress on May 14, 2019 8:00 PM EST

Gate

Planar FET

FinFET

ANANDTECH

PC COMPONENTS SMARTPHONES & TABLETS SYSTEMS ENTERPRISE

TRENDING TOPICS CPUS INTEL AMD MOBILE SMARTPHONES STORAGE RYZEN SSDS GPUS

Home > CPUs

Intel Details Manufacturing through 2023: 7nm, 7+, 7++, with Next Gen Packaging

237

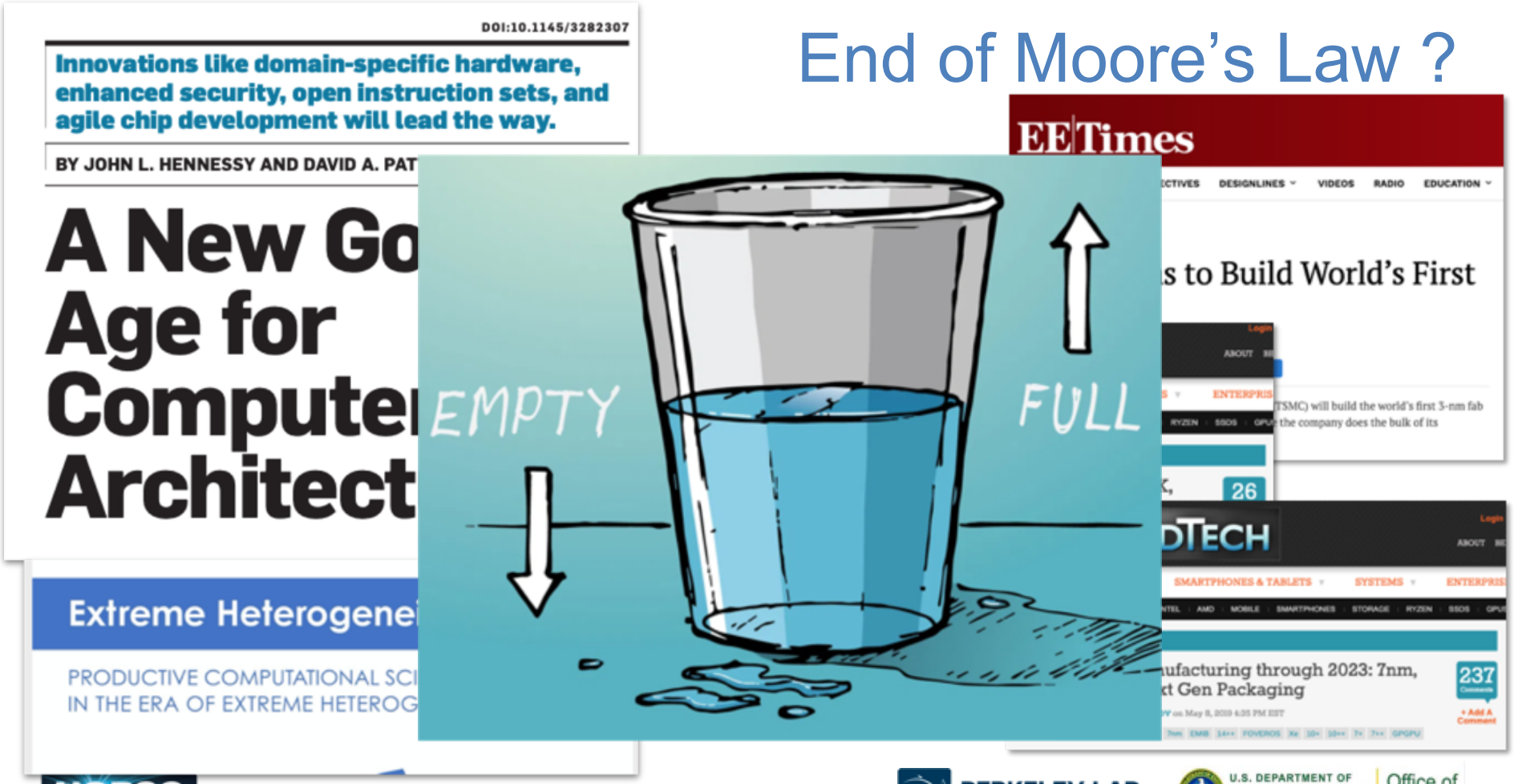
by Ian Cutress & Anton Shilov on May 8, 2019 4:35 PM EST

NERSC

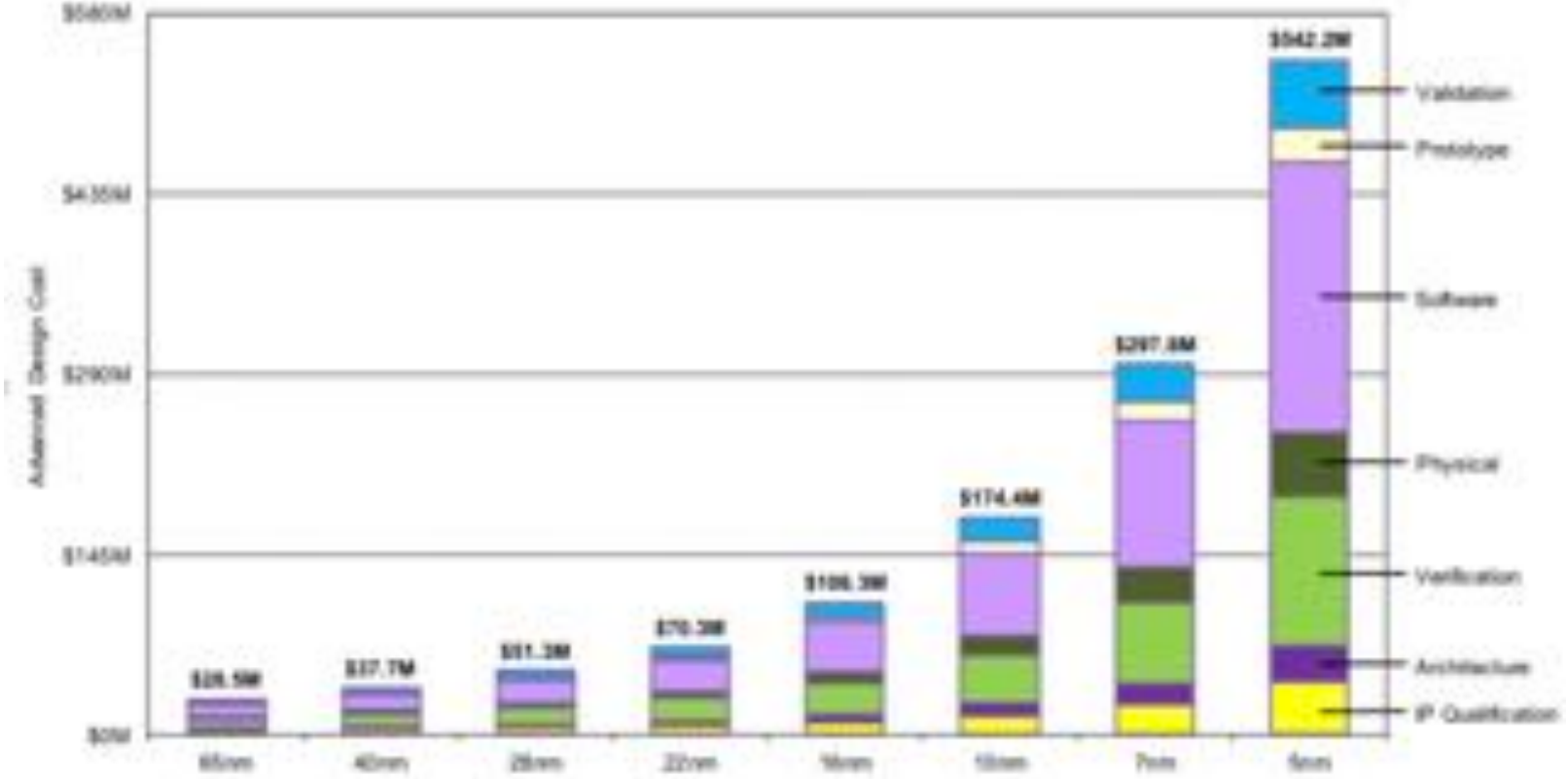
BERKELEY LAB Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY Office of Science

End of Moore's Law ?



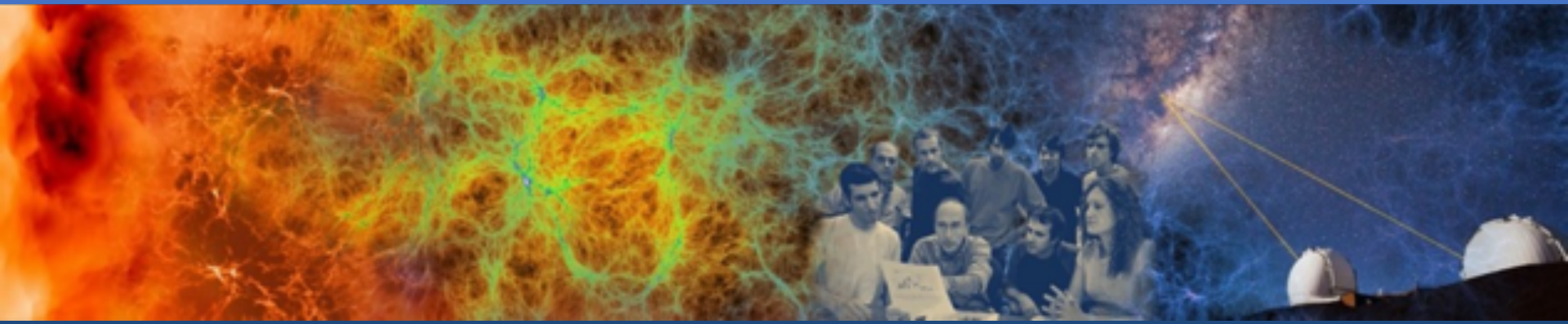
Can I just do it myself ? - Chip Design Costs



Summary

- Future HPC resources are likely to contain CPUs & GPUs
  - They continue to increase in performance each generation
    - Almost certainly on a per $ basis
    - Maybe not as much per Watt ~1.5x per generation
  - Different sites may have continue to have differing amounts of each - although more tightly-coupled solutions may affect this
- Lifetime of systems will increase as incentive to upgrade gets less
  - Expect more specialization

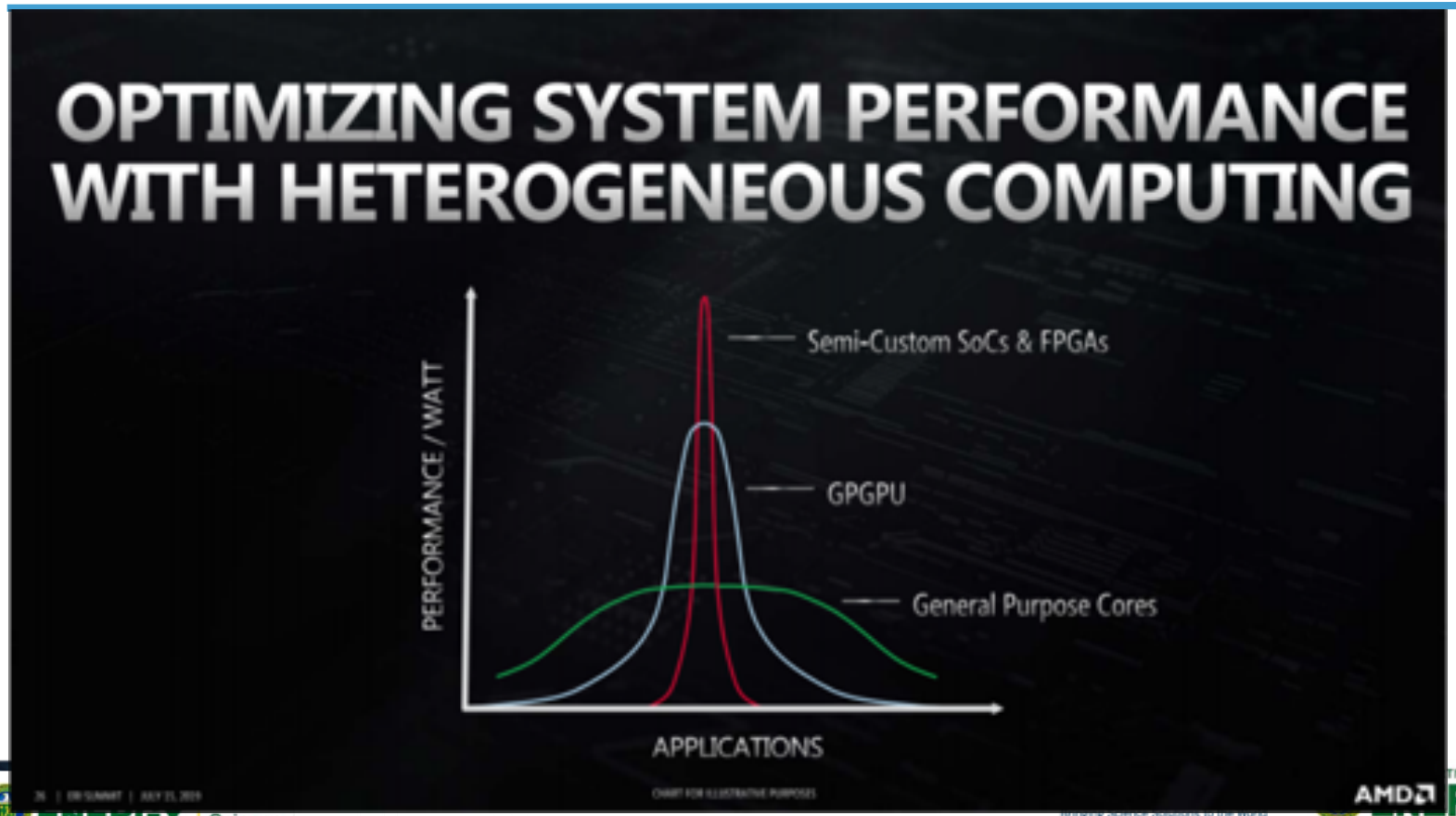
Questions ?



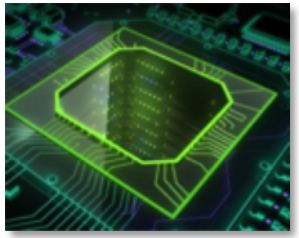
Will GPUs work for everybody?

- Will 100% of the NERSC workload be able to utilize GPUs by 2024?
  - Yes, they just need to modify their code
  - No, their algorithm needs changing
  - No, their physics is fundamentally not amenable to data parallelism
  - No, they just don't have time or need too

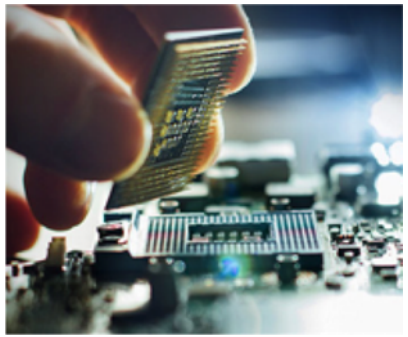
View from AMD - can we exploit this to benefit NERSC users ?



Specialization: End Game for Moore's Law



NVIDIA builds deep learning appliance with V100 Tesla's



Intel buys deep learning startup, Nervana

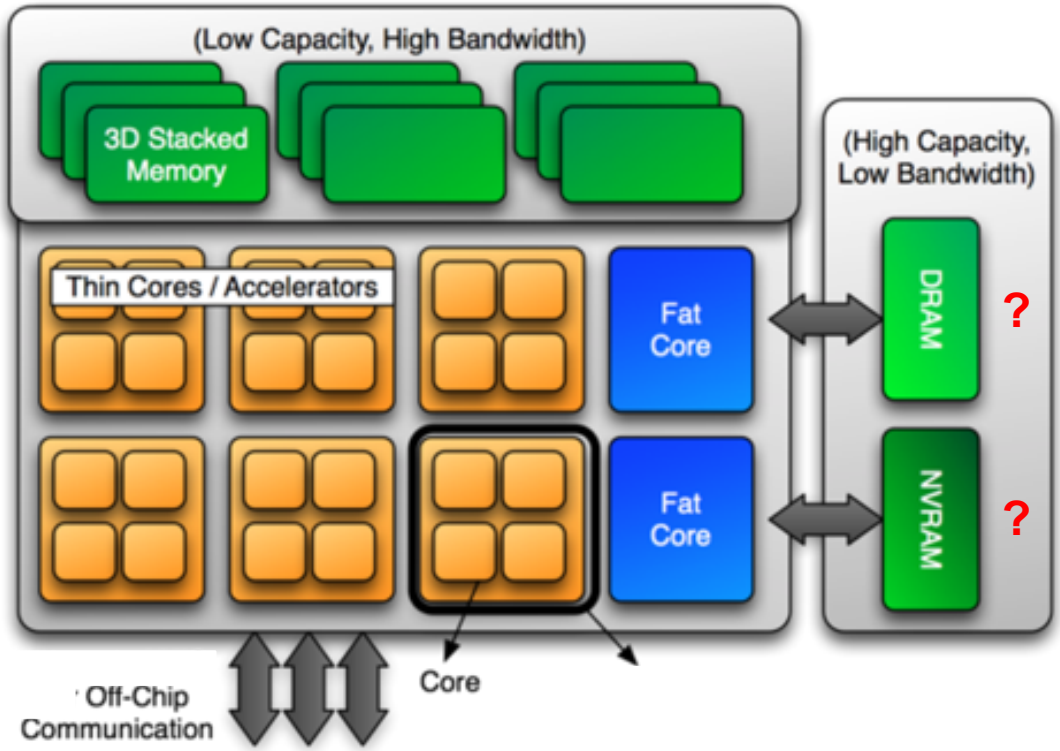
FPGAs offer configurable specialization

RISC-V

RISC-V is an open hardware platform

Google designs its own Tensor Processing Unit (TPU)

Potential 2024 Node



- Vendors converging to a mixture of energy-efficient Thin Cores/Accelerators and Fat Cores
- Potentially with DRAM/NVRAM
- (Hopefully) leads to less focus on data motion and more on identifying parallelism

J. A. Ang *et al.*, "Abstract Machine Models and Proxy Architectures for Exascale Computing," *2014 Hardware-Software Co-Design for High Performance Computing*, New Orleans, LA, 2014, pp. 25-32.
doi: 10.1109/Co-HPC.2014.4 http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7017960&isnumber=7017953

21